Greg Damico

Instructional Coordinator, Data Science

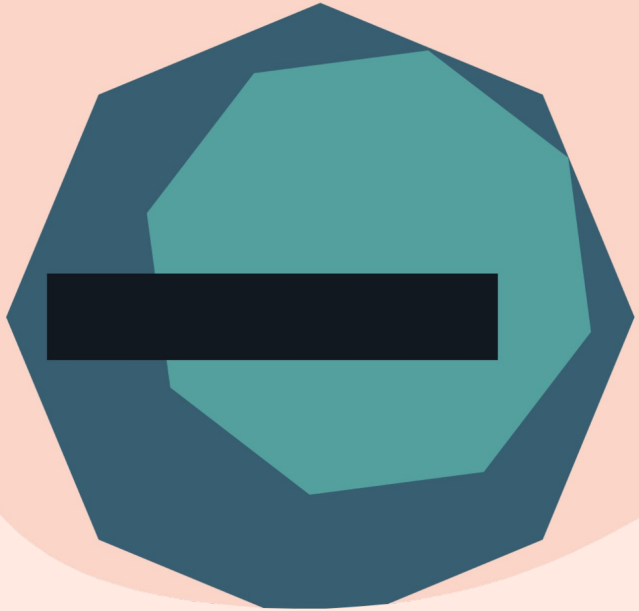
// FLATIRON SCHOOL

Bag-of-Words Models

What are they and how do they work?

// FLATIRON SCHOOL

Task: Make use of natural language as input to a machine learning model.

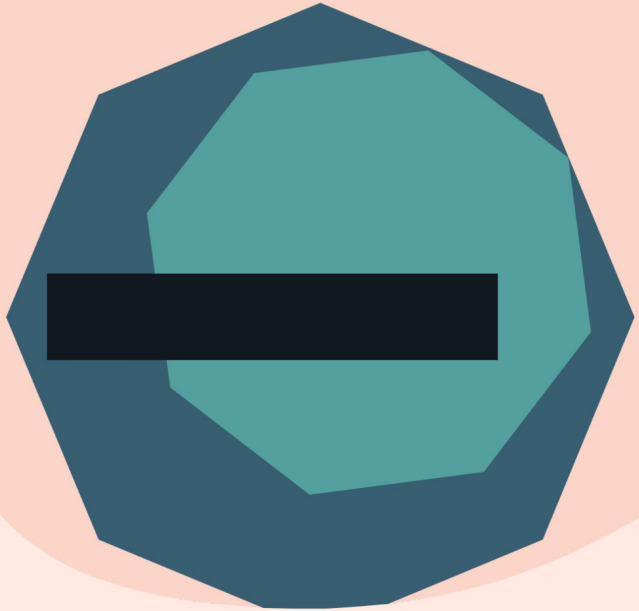


Possible Use Cases:

- Authorship Recognition
- Genre Classification
- SEO
- Spam Filtering

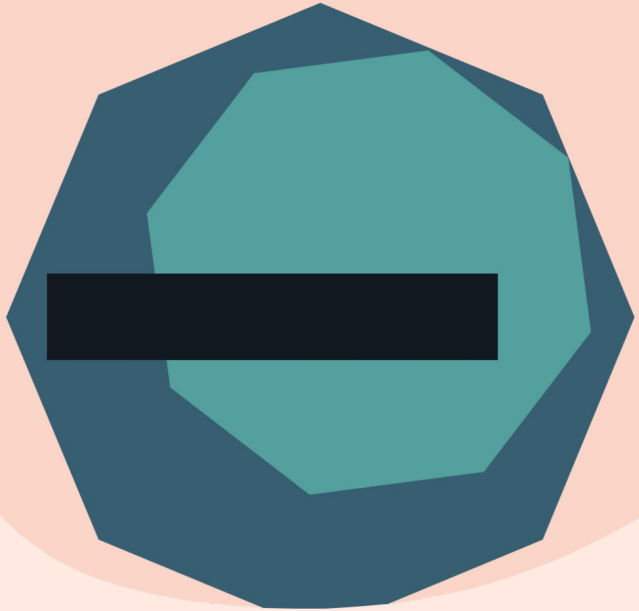
Assumptions:

- There is a **corpus** of **documents**
- Each document comprises a list of **tokens**



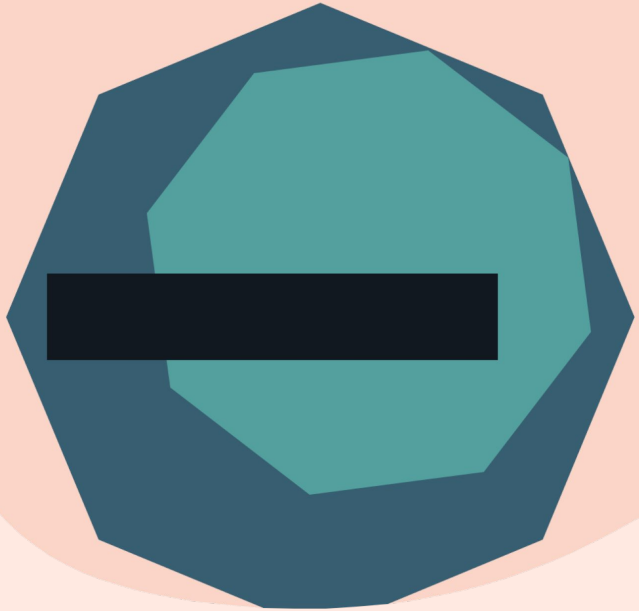
Two Main Ideas

- **Construct** a Vocabulary
- **Score** documents according to that Vocabulary



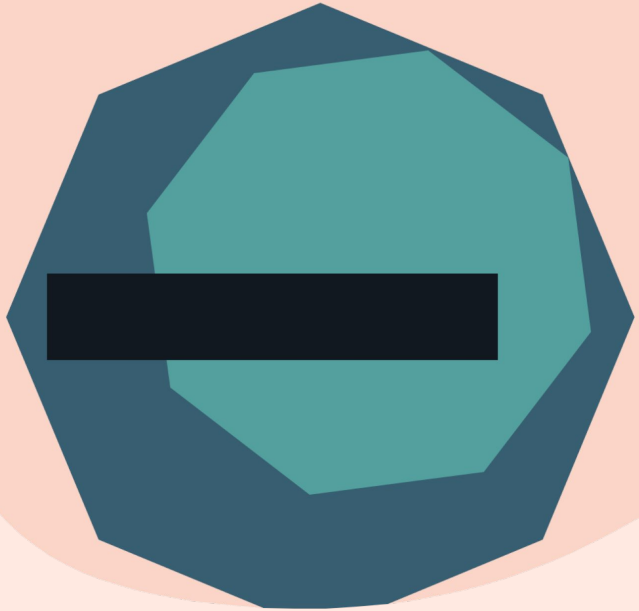
Vocabulary Construction

- Simplest Idea: Set of all words from all documents in corpus
- “Bag”: No structure, no order



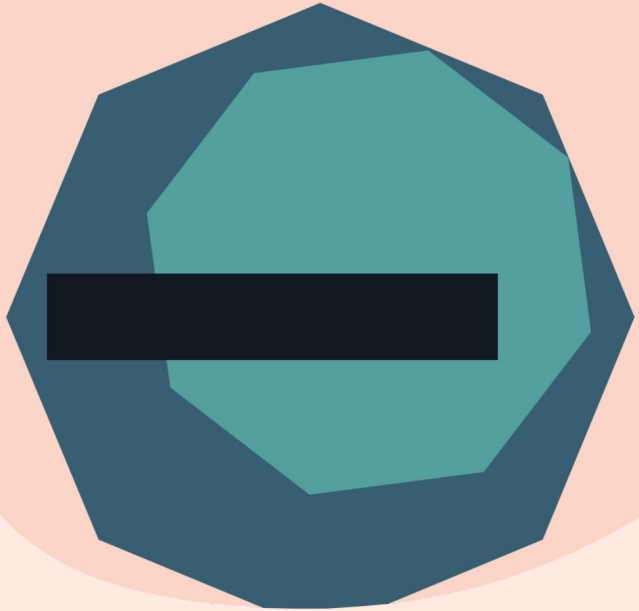
Vocabulary Construction

- Standard Steps:
  - Remove capitalization
  - Remove punctuation
  - Remove stopwords
  - Use stems / lemmas



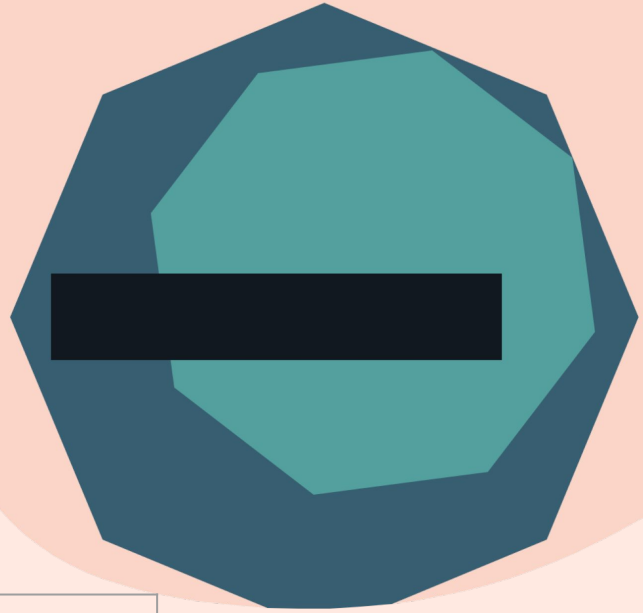
Vocabulary Construction

- More complexity:
  - Customize list of stopwords
  - Add ***N*-grams**: groups of *N* consecutive words



Document Scoring

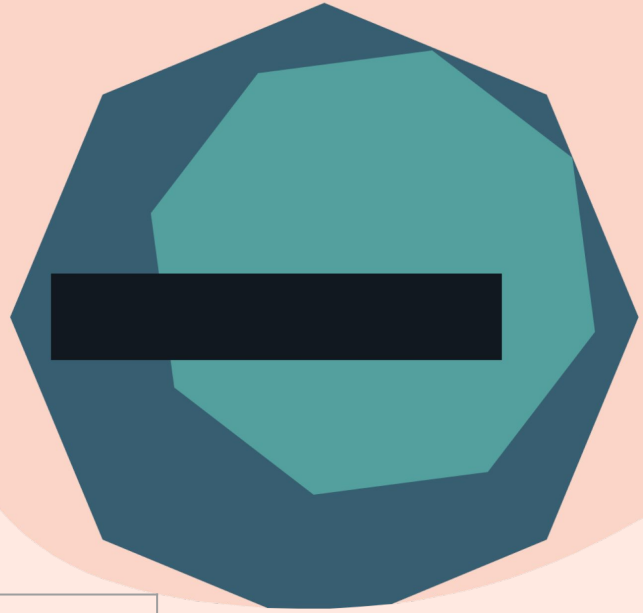
- Simplest Idea is Boolean: Token present in document or not



Document	*street*	*listen*	*go*	*outside*
doc1	1	0	1	1
doc2	1	1	0	0

Document Scoring

- More complexity:
    - Count Vectorization
    - TF-IDF Vectorization



Document	*street*	*listen*	*go*	*outside*
doc1	10	0	13	5
doc2	4	19	0	0